DOCUMENT RESUME

ED 227 137 TM 830 131

AUTHOR Schmitt, Neal
TITLE Formula Estimation of Cross-Validated Multiple Correlation.
PUB DATE Aug 82
NOTE 13p.; Paper presented at the Annual Meeting of the American Psychological Association (90th, Washington, DC, August the functional 23-27, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; *Estimation (Mathematics); *Mathematical Formulas; Mathematical Models; *Multiple Regression Analysis; Predictive Measurement; Research Methodology; Weighted Scores
IDENTIFIERS *Cross Validation; Predictive Models

ABSTRACT

A review of cross-validation shrinkage formulas is presented which focuses on the theoretical and practical problems in the use of various formulas. Practical guidelines for use of both formulas and empirical cross-validation are provided. A comparison of results using these formulas in a range of situations is then presented. The result of these comparisons indicate that one should use Cattin's formula to estimate cross-validated R, employing either Wherry or Olkin-Pratt estimates of the population R. If examination of predictor-criterion correlations has occurred prior to regression analysis, use empirical cross-validation, or adjust p to indicate the original number of variables examined. Double cross-validation is considered inefficient and unsatisfactory, and a cautionary remark concerning the functional number of predictors is presented. (Author/PN)

ED227137

Formula Estimation of Cross-Validated

Multiple Correlation

Neal Schmitt

Michigan State University

Running Head: Estimates of Cross-Validated Multiple Correlation

Send Correspondence to:

Neal Schmitt
Department of Psychology
Psychology Research Building
Michigan State University
E. Lansing, MI 48824-1117

TM 830 131

Abstract

A review of cross-validation shrinkage formulas is presented which focuses on the theoretical and practical problems in the use of various formulas. A comparison of results using these formulas in a range of situations is then presented. The result of these comparisons is that use of Cattin's formula is recommended. Double cross-validation is considered inefficient and unsatisfactory and a cautionary remark concerning the functional number of predictors is presented.

Formula Estimation of Cross-Validated Multiple Correlation

In 1931, Wherry published a formula subsequently used to estimate the multiple correlation between actual measures on some criterion variable and predicted values of that same variable. The predictions, of course, are made using regression weights developed in a sample from the same population concerning which the predictions are made. Wherry himself recognized that his formula really was an estimate of what the multiple correlations between predicted and actual criterion values would be if one had the population or true regression weights instead of those derived from some fallible sample. Because of the recognition that the formula was conceptually inappropriate (Wherry, 1951) and because of bad experiences with applications of the formula (Guion, 1965), most authors concerned with the stability of their prediction equations used actual empirical cross-validation of the type described by Mosier (1951). Briefly, Mosier proposed splitting the sample in half, computing regression equations and associated multiple correlations in both halves, and then using the regression equations developed in one half to make predictions about values of the criterion in the other half. Correlations between actual and predicted values for these two cross-validations were averaged to provide an estimate of the cross-validity. Mosier's procedure was called double cross-validation.

In 1977, Schmitt, Coyle, and Rauschenberger evaluated the performance of the Wherry estimate and two other similar formulas (Darlington, 1968; Nicholson, 1960) and the double cross-validation technique. The evaluation of these four methods was done in a Monte Carlo study using (1) the difference between the estimated cross-validated R and the actual population cross-validity and (2) the standard deviation of these estimates. Several guidelines for the

usage of these formulas were presented, most significantly that actual empirical cross-validation was inefficient and likely to be in greater error in any single application than any of the formulas including the Wherry formula.

Since that time several papers have appeared which have raised issue with the appropriateness of the formulas evaluated by Schmitt et al. (1977). Rozeboom (1978) indicated their conceptual inadequacy and Rozeboom (1978) as well as Drasgow, Dorans, and Tucker (1979) have shown that for low levels of multiple correlation not sampled by Schmitt et al. (1977), the formulas, particularly Darlington's, produced a severe negative bias. That is, for low levels of sample multiple correlation, the formula estimates of cross-validated multiple correlation were muoh too low.

Since that time there has also been general agreement (Cattin, 1980a; 1980b; Rozeboom, 1978; 1981) that a fourth formula presented by Browne (1975) is mathematically correct. Table 1 is a presentation of various formula estimates.

Insert Table 1 about here

As can be seen, the Browne formula is horrendous from a computational viewpoint. Hence recent efforts (Cattin, 1980a; Rozeboom, 1981) have focused on the development and evaluation of shortcut formulas which yield essentially the same values as the formula presented by Browne (1975).

The purpose of this paper is to present briefly some comparisons of these formulas, parts of which are available in the citations listed above. Second, I will attempt to provide practical guidelines for use of both formulas and empirical cross-validation.

Method

A range of possible sample squared multiple correlations (.1 to .9), sample sizes (40 to 240) and number of predictors (5 to 25) was selected to be reasonably representative of applied research employing multiple regression.

The various formulas presented in Table 1 were then applied to these sample statistics to provide estimates of the population corss-validity in any given situation.

Results and Discussion

In Table 2, are cross-validity estimates based on the various proposed formulas (see Table 1). Various levels of sample multiple correlation (R),

----
Insert Table 2 about here
----

sample size (N), and number of predictors (P), are used in these computations. If one examines Table 2, it becomes obvious that for relatively large N/P ratios there are larger differences, as various authors have pointed out (Rozeboom, 1978; Drasgow, Dorans, & Tucker, 1979), and they are likely practically important differences.

The other factor that is extremely important practically is that the Nicholson and Darlington formula fail for low levels of multiple correlation ($R^2 < .6$) which is precisely the levels of multiple correlation typically found in applied situations. This failure, of course, is the one noted by various authors cited above. Finally, even the Cattin and Rozeboom alternatives produce impossible results when the N/P ratio and R^2 is small. The underlined values in Table 2 are illustrative of this problem.

The most significant conclusion to be drawn from these results as well as the other cited literature on this topic is that Cattin's formula is the most appropriate estimate of the cross-validated multiple correlation. Note that Cattin's formula requires the use of Wherry's formula to calculate the population multiple correlation. Further, it seems appropriate that use of even their formula be restricted to instances in which N/P is greater than 2 especially when R^2 is low ($<.6$).

At least one other practical issue remains. Is it better to use the Cattin formula or empirical cross-validation? My answer is that empirical cross-

validation is not only a waste of time, it is is less satisfactory than any formula estimate. The reason for this was displayed in Table 3 of Schmitt, Coyle, and Rauschenberger (1977). Empirical cross-validation, since it is based on substantially less than the total sample, is associated with greater variance across replications than are formula estimates. So, in any given instance, we can be much more wrong in our estimate with empirical cross-validation than if we had applied one of the formulas available.

A final note of caution in the use of formula estimates of cross-validation is that they assume there has been no "data-snooping" prior to the calculation of the sample regression equation. The procedure in some studies is to compute zero-order correlations between a criterion and a large number of predictors, pick those variables which are significantly related to the criterion and compute a regression equation and multiple correlation using this subset of significant predictors. The functional p in this case is not the number of predictors in the regression equation but the total number of potential predictors for which correlations were observed.

Conclusions

The conclusions are simple: 1) use the Cattin formula to estimate cross-validated R employing either Wherry or Olkin-Pratt estimates of the population R (see Cattin, 1980a for details); and 2) if examination of predictor-criterion correlations has occurred prior to regression analysis, use empirical cross-validation or adjust p to indicate the original number of variables examined.

References

Browne, M.W. A comparison of single sample and cross-validation methods for estimating the mean squared error of prediction in multiple linear regression. British Journal of Mathematical and Statistical Psychology, 1975, 28, 112-120.

Cattin, P. Note on the estimation of the squared cross-validated multiple correlation of a regression model. Psychological Bulletin, 1980, 87, 63-65. (a)

Cattin, P. Estimation of the predictive power of a regression model. Journal of Applied Psychology, 1980, 65, 407-414. (b)

Darlington, R.B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Drasgow, F., Dorans, N.J., & Tucker, L.R. Estimates of the squared cross-validity coefficient: A Monte Carlo investigation. Applied Psychological Measurement, 1979, 3, 387-399.

Mosier, C.I. I. Problems and designs of cross-validation. In Symposium: The need and means of cross-validation. Educational and Psychological Measurement, 1951, XI, 5-11.

Nicholson, G.E. Prediction in future samples. In I Olkin et al. (Eds.), Contributions to probability and statistics. Stanford, CA: Stanford University Press, 1960.

Olkin, E., & Pratt, J.W. Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics, 1958, 29, 201-211.

Rozeboom, W.W. The estimation of cross-validated multiple correlation: A clarification. Psychological Bulletin, 1978, 85, 1348-1351.

Rozeboom, W.W. The cross-validational accuracy of sample regressions. Journal of Educational Statistics, 1981, 6, 179-198.

Schmitt, N., Coyle, B.W., & Rauschenberger, J.A. Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. *Psychological Bulletin*, 1977, *84*, 751-758.

Wherry, R.J., Sr. A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, 1931, *2*, 440-457.

Wherry, R.J., Sr. IV. Comparison of cross-validation with statistical inference of betas and multiple R from a single sample. In Symposium: The need and means of cross-validation. *Educational and Psychological Measurement*, 1951, *XI*, 23-28.

Table 1

Summary of Cross-Validation Formula

Author	Formula	Comment
Wherry (1931)	$\rho c = 1 - \left(\frac{N-1}{N-p-1}\right)(1-R^2)$	Estimates multiple R when we have population weights
Nicholson (1960)	$\hat{\rho}_c^{\,2} = 1 - \left(\frac{N-1}{N-p-1}\right)\left(\frac{N+p+1}{N}\right)(1-R^2)$	Developed for fixed and random effects models respectively - both suffer negative bias $\geq$ .1 when N/P < 2.
Darlington (1968)	$\hat{\rho}_c^{\,2} = 1 - \left(\frac{N-1}{N-p-1}\right)\left(\frac{N-2}{N-p-2}\right)\left(\frac{N+1}{N}\right)(1-R^2)$	
Rozeboom (1978)	$\hat{\rho}_c^{\,2} = \rho^2\left[1+\left(\frac{p}{N-p-2}\right)\left(\frac{1-\rho^2}{\rho^2}\right)\right]^{-1}$	
Cattin (1980a, 1980b)	$\hat{\rho}_c^{\,2} = \frac{(N-p-3)\,\rho^4 + \rho^2}{(N-2p-2)\,\rho^2 + p}$	First portion of Browne formula (1975). ρ must be estimated separately by a formula such as Wherry's above or in cases of very low N, a formula provided by Olkin and Pratt (1958).
Brown (1975)	$\hat{\rho}_c^{\,2} = \left[\frac{(N-p-3)\,\rho^4 + \rho^2}{(N-2p-2)\,\rho^2 + p}\right] - \frac{2(N-p-2)\,(N-2p-6)\,(p-1)\,\rho^4\,(1-\rho^2)^2}{(N-p-4)\,\left[(N-2p-2)\,\rho^2 + p\right]^3} + 0\left[(N-p)^{-1}\right]$	

[a] In all formulas, R = sample multiple correlation, N = sample size, p = number of predictor variables, ρ = population multiple correlation, $\hat{\rho}_c$ = population cross-validity.

Table 2

Estimates of ρ_c^2 Based on Wherry, Nicholson, Rozeboom Darlington, and Cattin Formulas for Various Combinations of R^2, N, and p

R^2	N	P	Wherry	Nicholson	Darlington	Rozeboom	Cattin
.9	40	5	.89	.87	.86	.87	.88
.9	80	5	.89	.89	.89	.89	.89
.9	240	5	.90	.90	.90	.90	.90
.9	40	10	.87	.83	.81	.82	.83
.9	80	10	.89	.87	.87	.87	.87
.9	240	10	.90	.89	.89	.89	.89
.9	40	25	.73	.54	.17	.41	.44
.9	80	25	.85	.81	.78	.79	.79
.9	240	25	.89	.88	.88	.88	.88
.8	40	5	.77	.74	.73	.74	.75
.8	80	5	.79	.77	.77	.77	.78
.8	240	5	.80	.79	.79	.79	.79
.8	40	10	.73	.66	.63	.65	.67
.8	80	10	.77	.74	.73	.74	.74
.8	240	10	.79	.78	.78	.78	.78
.8	40	25	.44	.08	-.67	.13	.15
.8	80	25	.71	.61	.56	.59	.60
.8	240	25	.78	.75	.75	.75	.75
.6	40	5	.54	.47	.46	.48	.51
.6	80	5	.57	.54	.54	.55	.55
.6	240	5	.59	.58	.58	.58	.59
.6	40	10	.46	.31	.25	.33	.36
.6	80	10	.54	.48	.47	.48	.49
.6	240	10	.58	.56	.56	.57	.57
.6	40	25	-.11	-.84	—[a]	.01	.00
.6	80	25	.42	.23	.13	.25	.26
.6	240	25	.55	.51	.50	.51	.51
.4	40	5	.31	.21	.19	.23	.26

Table 2 Continued

R^2	N	P	Wherry	Nicholson	Darlington	Rozeboom	Cattin
.4	80	5	.36	.31	.31	.32	.33
.4	240	5	.39	.37	.37	.37	.38
.4	40	10	.19	-.03	-.12	.08	.10
.4	80	10	.31	.22	.20	.24	.24
.4	240	10	.37	.35	.34	.35	.35
.4	40	25	-.07	---	---	.18	.17
.4	80	25	.12	-.16	-.31	.03	.03
.4	240	25	.33	.26	.25	.27	.27
.2	40	5	.08	-.06	-.08	.03	.04
.2	80	5	.15	.08	.08	.10	.11
.2	240	5	.18	.16	.16	.17	.17
.2	40	10	-.08	-.37	-.50	.02	.01
.2	80	10	.08	-.04	-.06	.03	.04
.2	240	10	.17	.13	.13	.14	.14
.2	40	25	---	---	---	.49	.51
.2	80	25	-.17	-.55	-.74	.08	.07
.2	240	25	.11	.10	-.02	.05	.06
.1	40	5	-.03	-.19	-.22	.01	.00
.1	80	5	.04	-.03	-.04	.02	.02
.1	240	5	.08	.06	.06	.07	.07
.1	40	10	-.21	-.54	-.68	.20	.19
.1	80	10	-.03	-.17	-.20	.01	.04
.1	240	10	.02	-.07	-.08	.00	.04
.1	40	25	---	---	---	.69	.73
.1	80	25	-.32	-.75	-.96	.33	.30
.1	240	25	-.01	-.11	-.13	.00	.00

[a]Values in cases with a blank were less than -1.00.